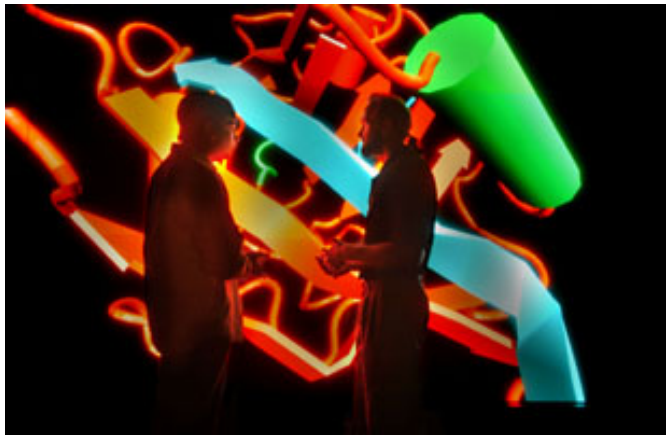
Enabling Next Generation Modeling and Simulations in Biology

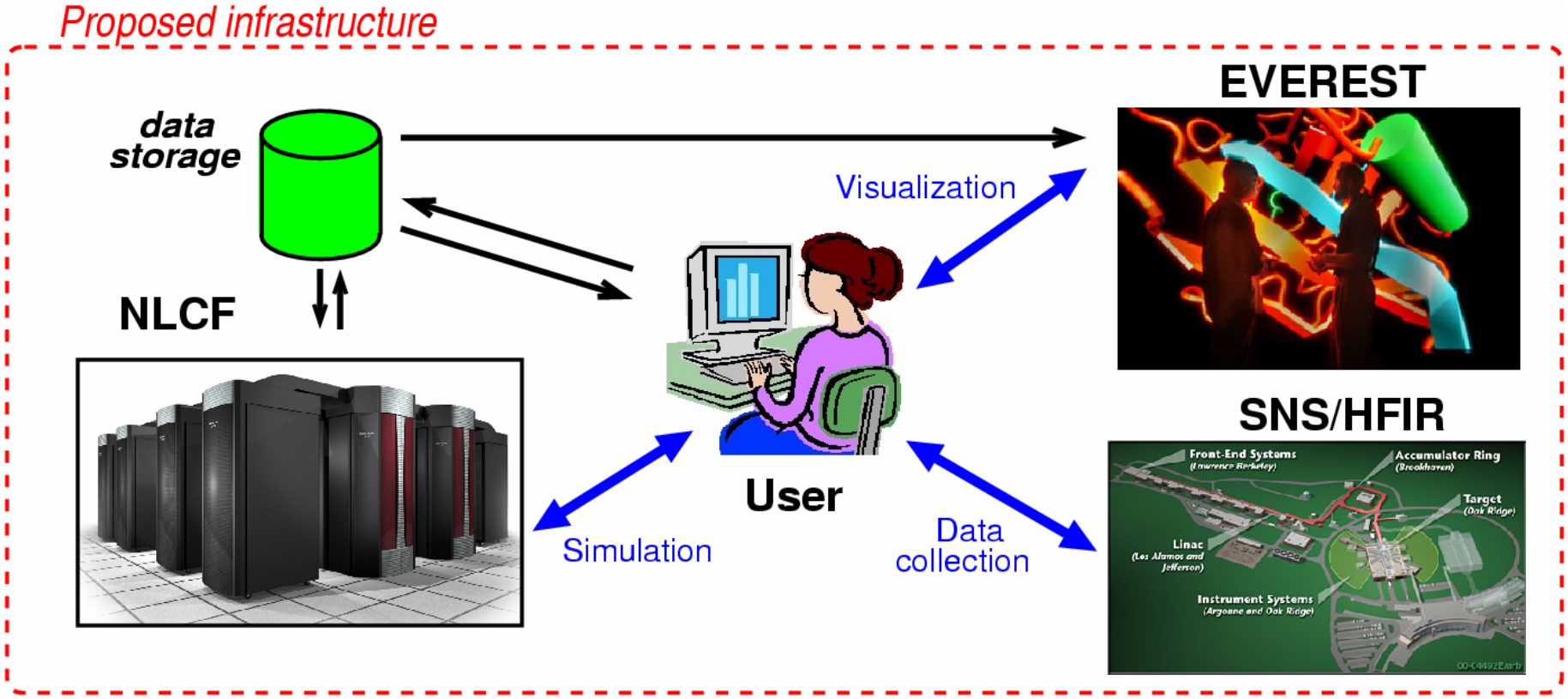


Pratul K. Agarwal, Sadaf R. Alam, Jeff S. Vetter (CSMD)

Ed Uberbacher (LSD) and Dean A. A. Myles (CSD)

Oak Ridge National Laboratory

Multi-disciplinary Infrastructure for Next Generation Modeling and Simulations



- Data collection from SNS/HFIR
- Simulations on NLCF machines with teraFLOP power
- Visualization on EVEREST and data storage on HPSS

Significance

- **computational genomics**
- **database design and enhancement**
- ***molecular modeling and simulation***
- **analysis of complex biological systems**
- **cell modeling**
- **biomedical software development**
- **high-throughput data analysis**

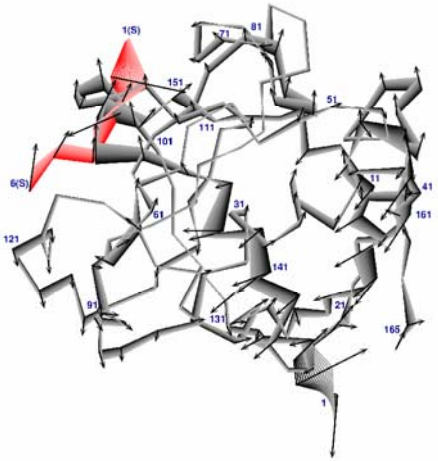
Detailed biophysical characterization of molecular machines, protein structure prediction, docking and others

Multi-scale modeling – Structure, Dynamics and Function

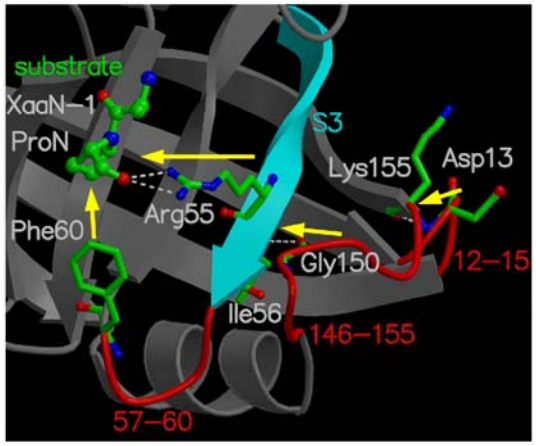
Desired/Current capability Ratio: 10^4-10^6

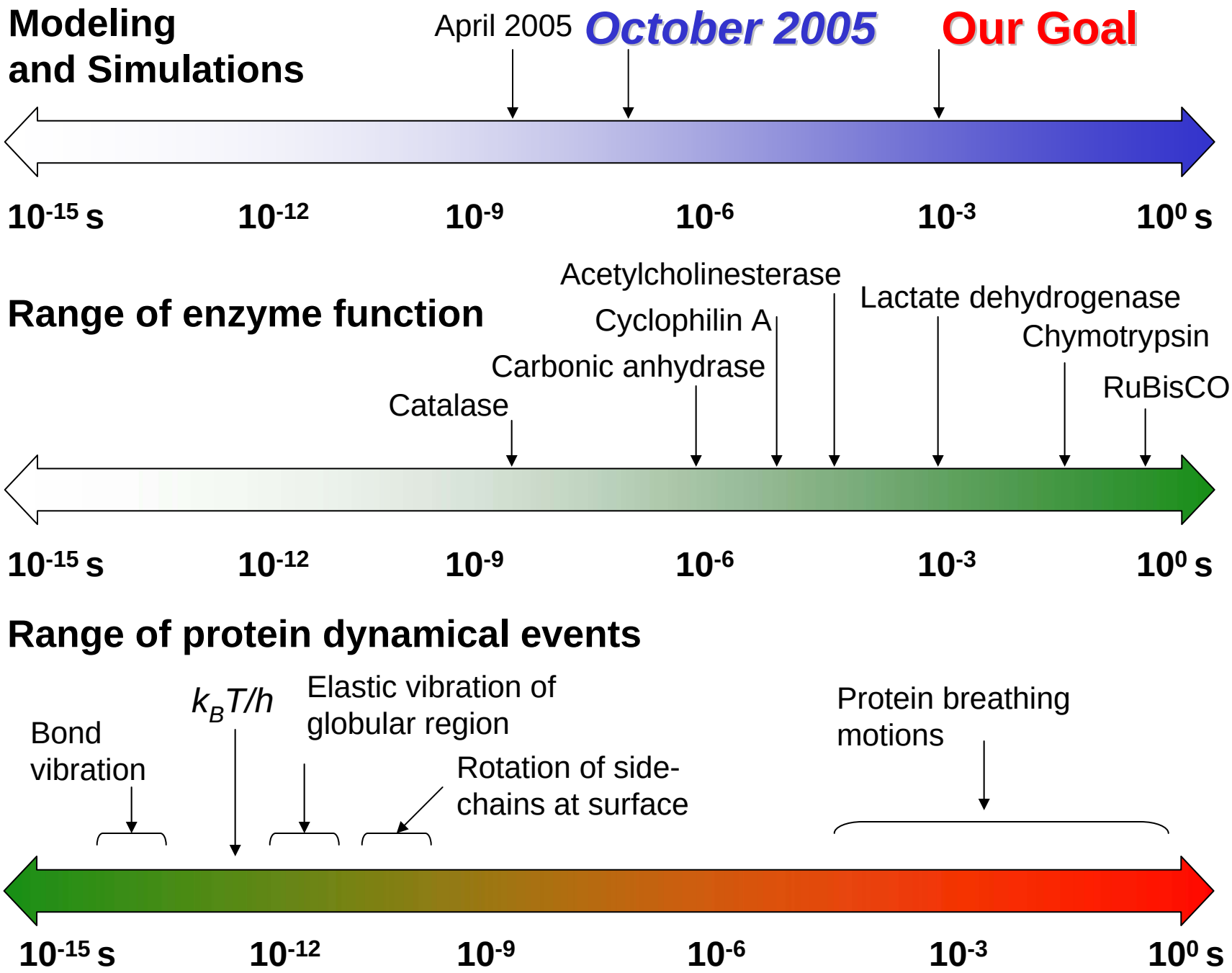
Structure

Dynamics



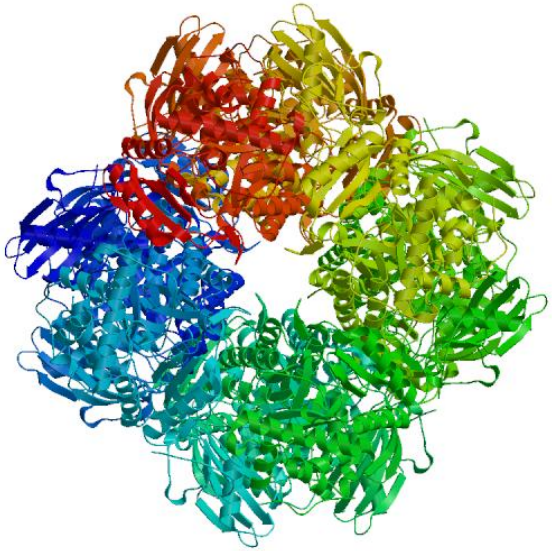
Function



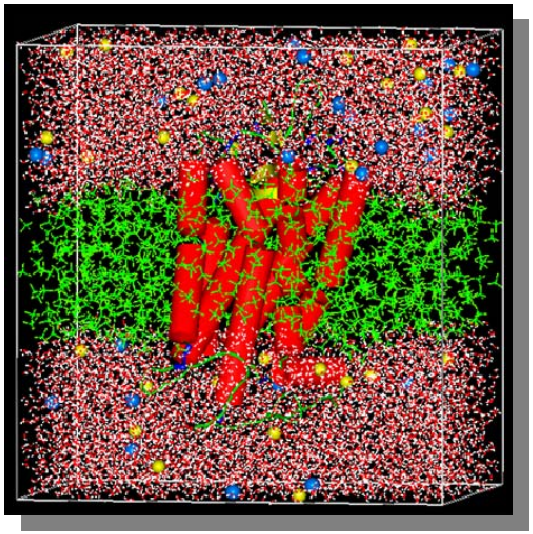

Modeling
and Simulations
April 2005
October 2005
Our Goal
10-15 s
10-12
10-9
10-6
10-3
100 s
Range of enzyme function
Acetylcholinesterase
Cyclophilin A
Lactate dehydrogenase
Chymotrypsin
Carbonic anhydrase
RuBisCO
Catalase
10-15 s
10-12
10-9
10-6
10-3
100 s
Range of protein dynamical events
Bond
vibration
kBT/h
Elastic vibration of
globular region
Rotation of side-
chains at surface
Protein breathing
motions
10-15 s
10-12
10-9
10-6
10-3
100 s

Enzyme RuBisCO

- Plays a role in Carbon sequestration
- The enzyme activity is at the order of (10^0) seconds
- Very large enzyme
- Multi-scale modeling needed

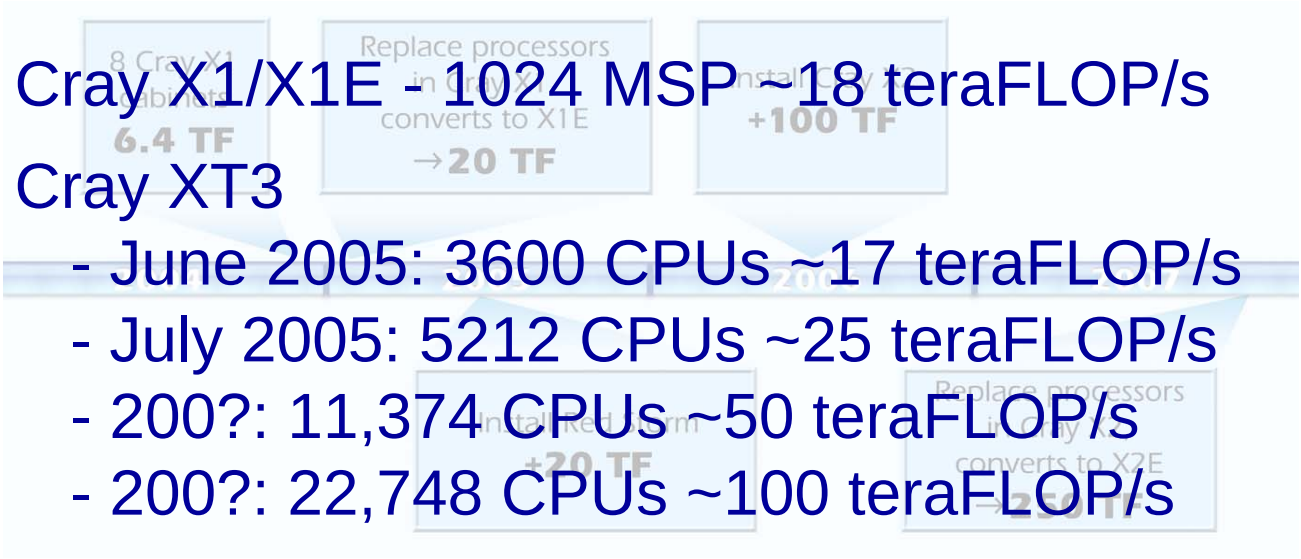


Biological Membranes



- Many cellular functions including transport, signaling and cellular interactions
- Computational modeling and simulations will have broad impact
- In combination with neutron and X-ray scattering, modeling and simulations will provide valuable insights

Porting and optimizing popular MM and QM codes in biology to NLCF machines



- Cray X1/X1E - 1024 MSP ~18 teraFLOP/s
- Cray XT3
  - June 2005: 3600 CPUs ~17 teraFLOP/s
  - July 2005: 5212 CPUs ~25 teraFLOP/s
  - 200?: 11,374 CPUs ~50 teraFLOP/s
  - 200?: 22,748 CPUs ~100 teraFLOP/s

- Popular molecular mechanics (MM) packages
  - AMBER, CHARMM, LAMMPS, NAMD, GROMACS
- Quantum-mechanical (QM) for detailed modeling
  - GAMESS, NWChem
- AMBER: Cheetah Supercomputer
  - speed up ~4 on 8 CPUs & ~8 on 32 CPUs
  - limit of maximum128 processors

Scaling the MD Kernel

$$V(\mathbf{r}^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2$$
$$+ \sum_{torsions} \frac{V_n}{2}(1 + cos(n\omega - \gamma))$$
$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)$$

Evaluations of non-bonded interactions most compute intensive part of MD: 80-90% of runtime!

Particle-mesh Ewald (PME) method

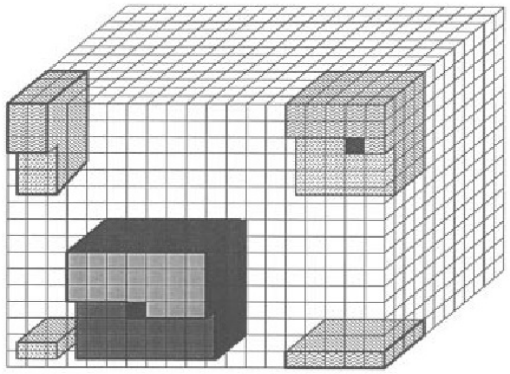
E(electrostatic)

= E(direct)

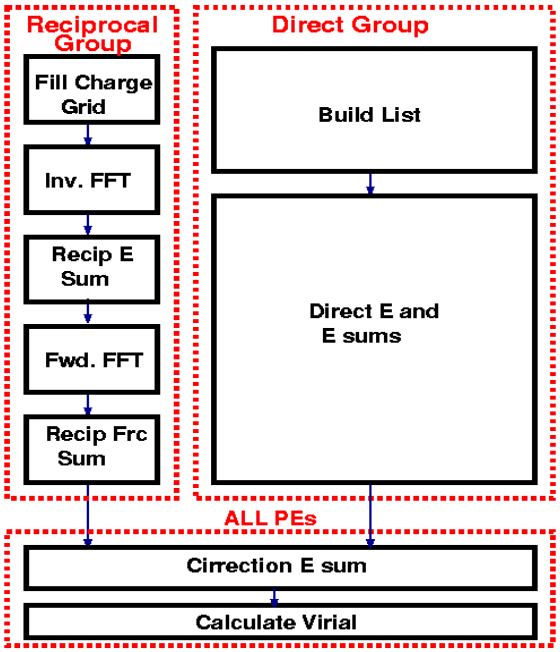
+ E(reciprocal)

+ E(correction)

Non-bonded pair list



Discrete Fourier transform on a grid



Preliminary Performance Analysis

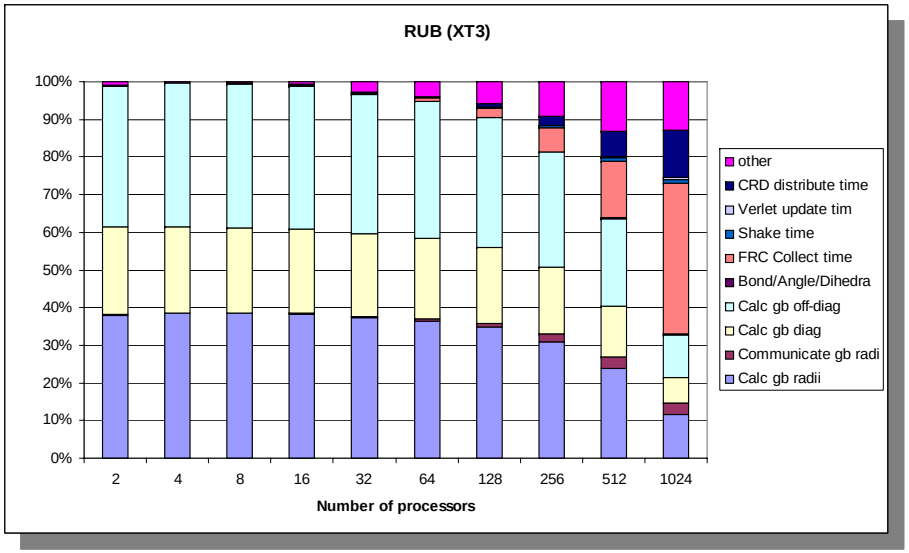
- AMBER scaled beyond 128 processors

  -1024 nodes on IBM BG/L

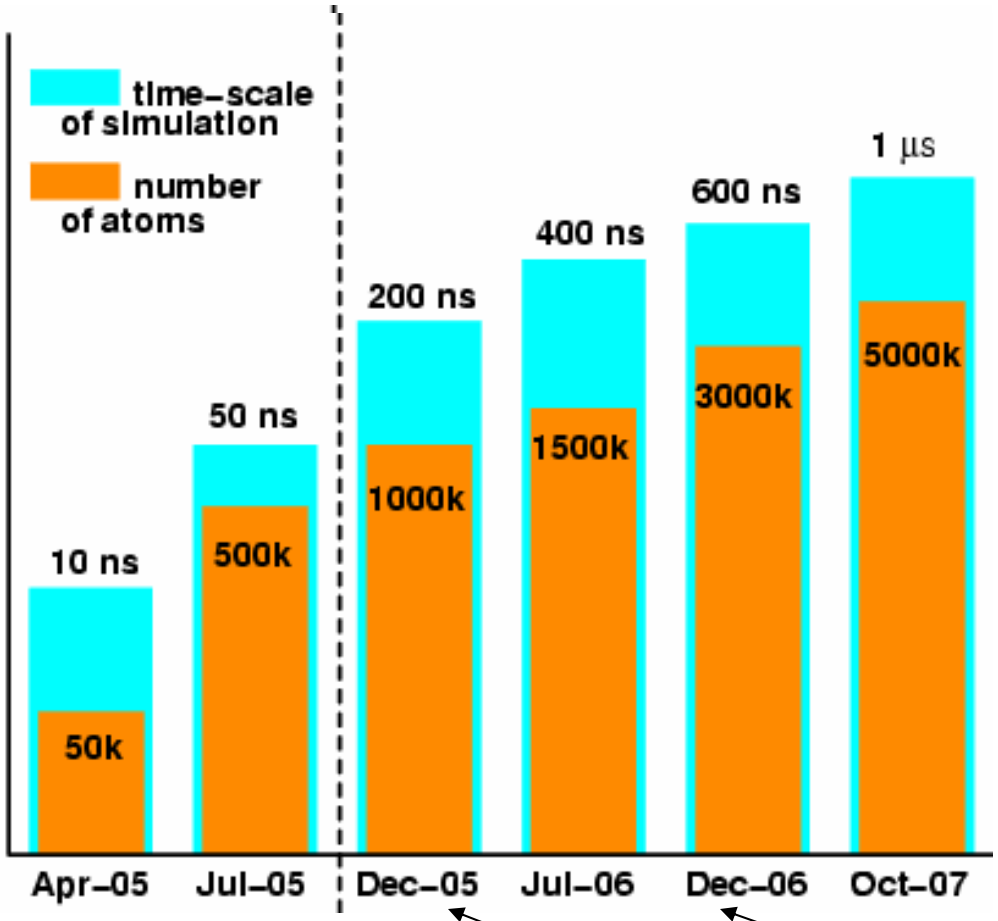
  - 4096 nodes on Cray XT3

- Application profiling: identified code blocks and communication patterns that limit the scalability

- LAMMPS – highly scalable MD engine



Time-line for scaling on NLCF machines



IBM p690
128 processors

ANL BGL
512-vn processors

ORNL XT3
1024 processors

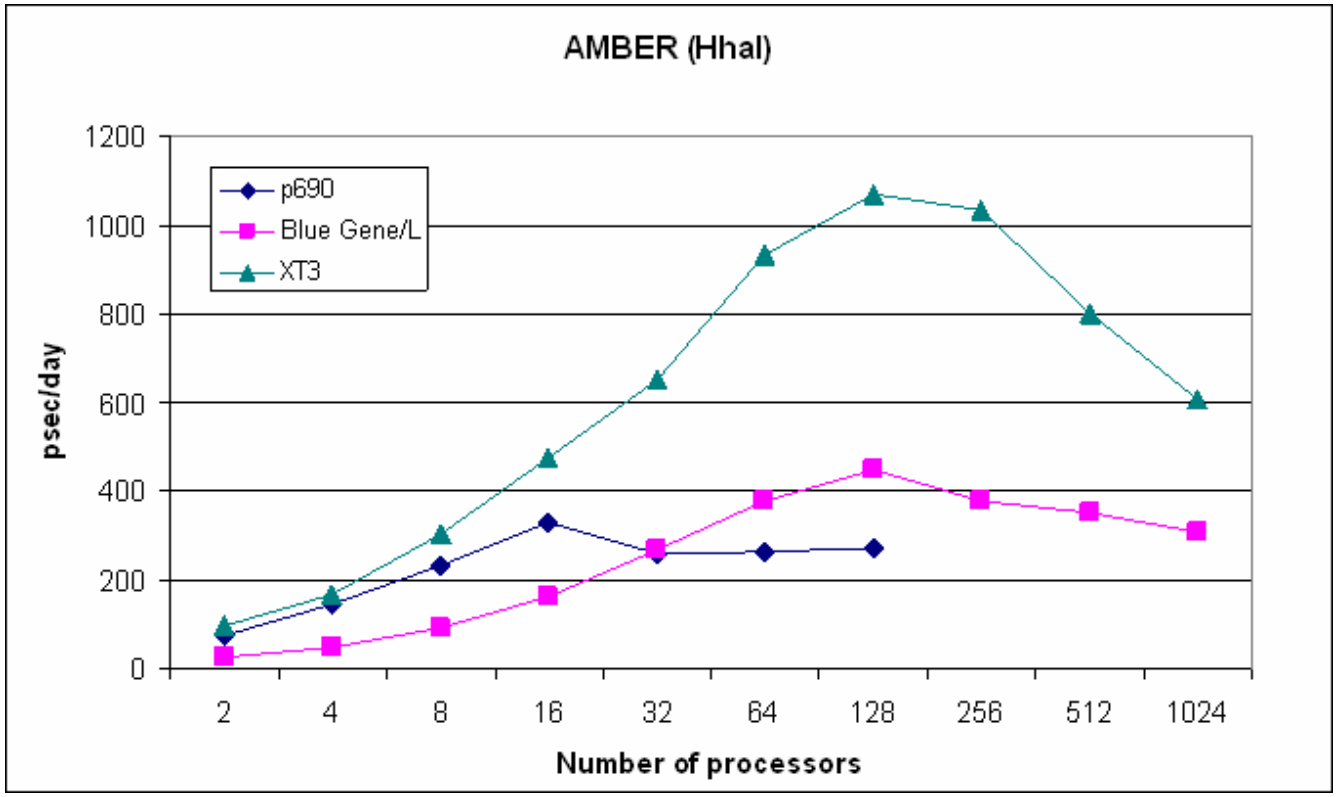
ORNL XT3
4096 processors

ORNL XT3
>16000 processors

200?
100 teraFLOP/s

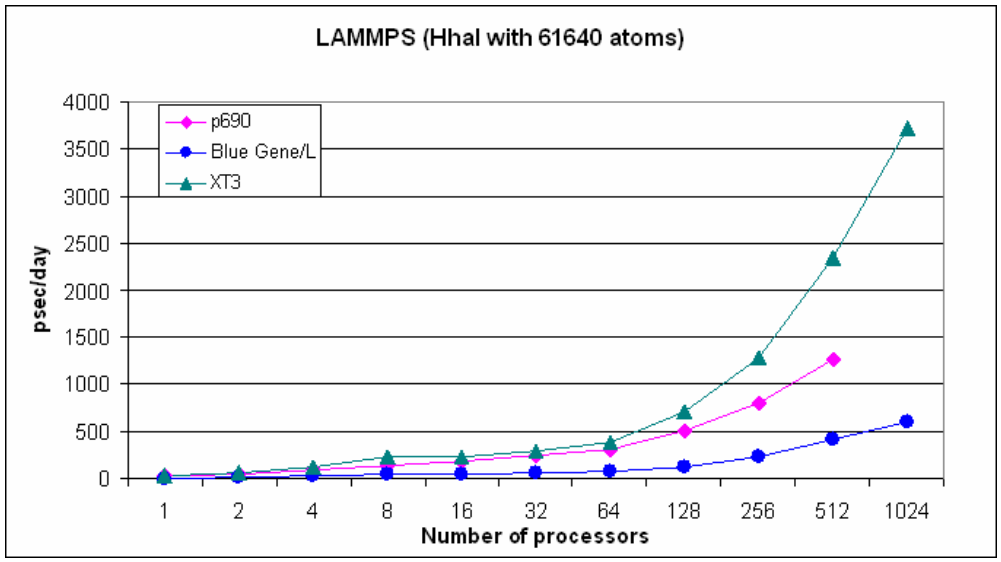
200?
>200 teraFLOP/s

Preliminary work: Scaling studies



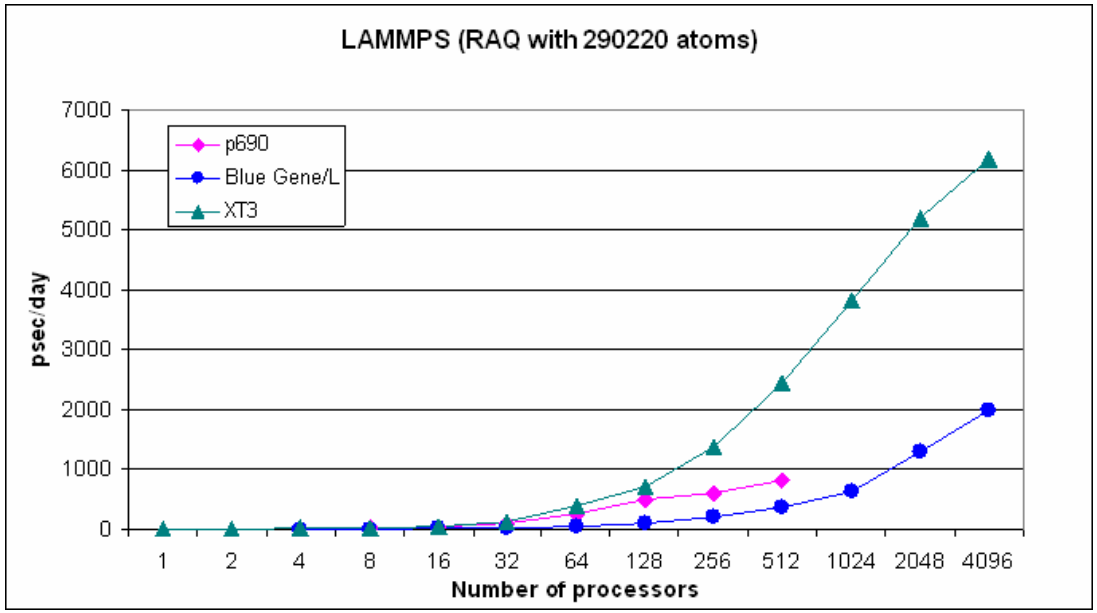
- Molecular dynamics: AMBER (v 8) *sander* module
- Solvated enzyme-DNA complex with about 60,000 atoms

Preliminary work: Scaling studies



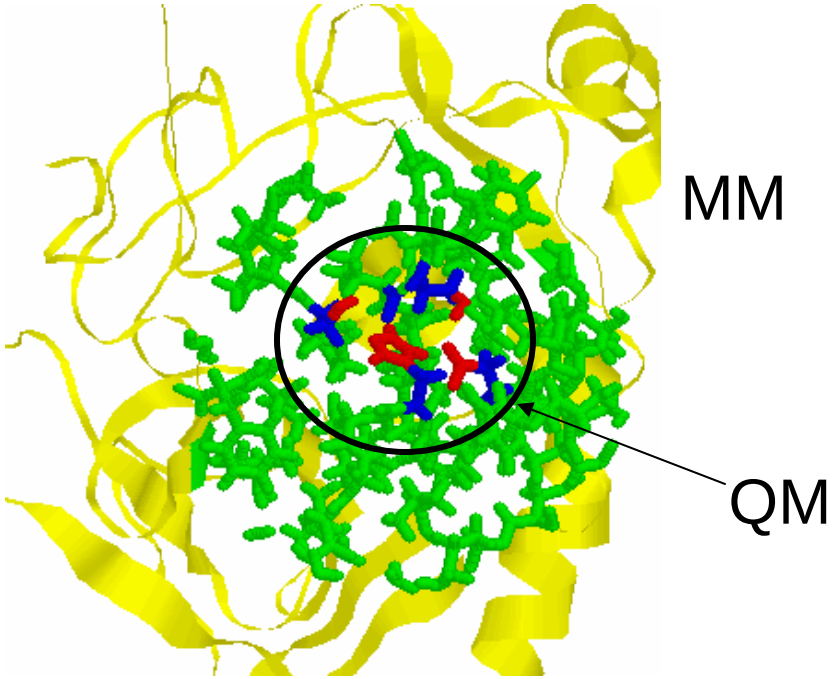
- LAMMPS with AMBER force-field
- Explicit solvent

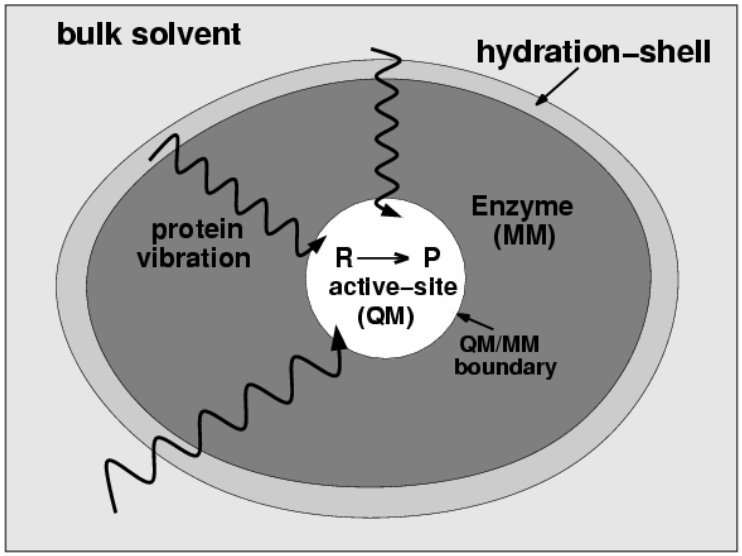
- 30 days - 200 ns!
- ~300K atoms!
- 16,000 processors on IBM Blue Gene/L



Hybrid QM/MM methodology for simulating enzyme catalysis

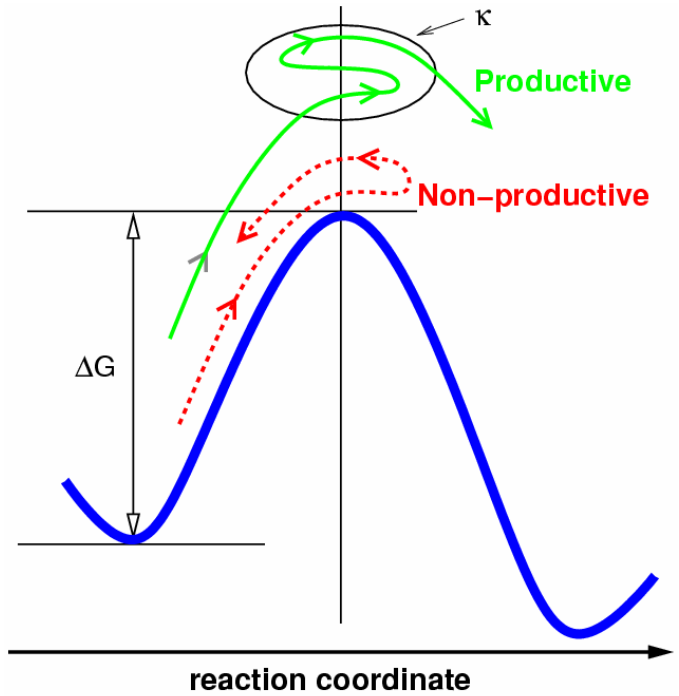
- Accurate treatment of long range electrostatics and dynamics
- Multi-scale protein dynamics: Novel ways to detect reaction coupled vibrational modes
- Explicit solvent: solvent-protein dynamics coupling





Multi-scale modeling

- Enzyme kinetics: Transition state theory
- Multi-scale protein dynamics: GNM, QHA, TANCA
- Fast time-scales (picosecond-microsecond)
- Mechanistic insights: local structure and long range effects
- The effect of dynamical effects on free energy, individual residues (mutants)
- Effect on barrier height and transmission coefficient
- Vibrational mode driven reactions

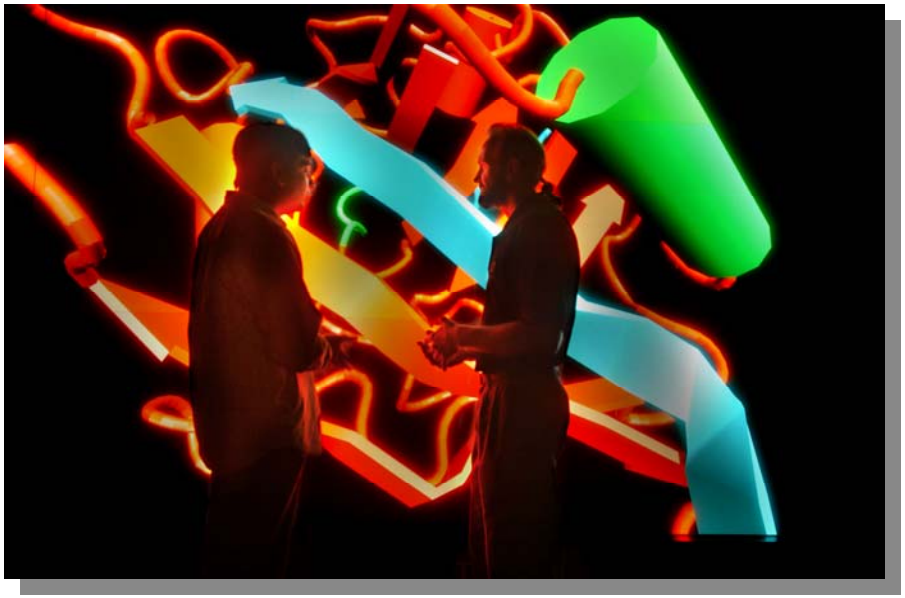


$$k_{TST} = \left(\frac{k_B T}{h}\right) \exp\left(\frac{-\Delta G^{\dagger}}{k_B T}\right)$$

$$k = \kappa k_{TST}$$

Interfacing with visualization facility and data storage

- Visualization of biological data is critical for fast interpretation and analysis of large data sets
- *Network of protein vibrations identified using EVEREST*
- Interface experimental and computational facilities with EVEREST: model low as well as high resolution data

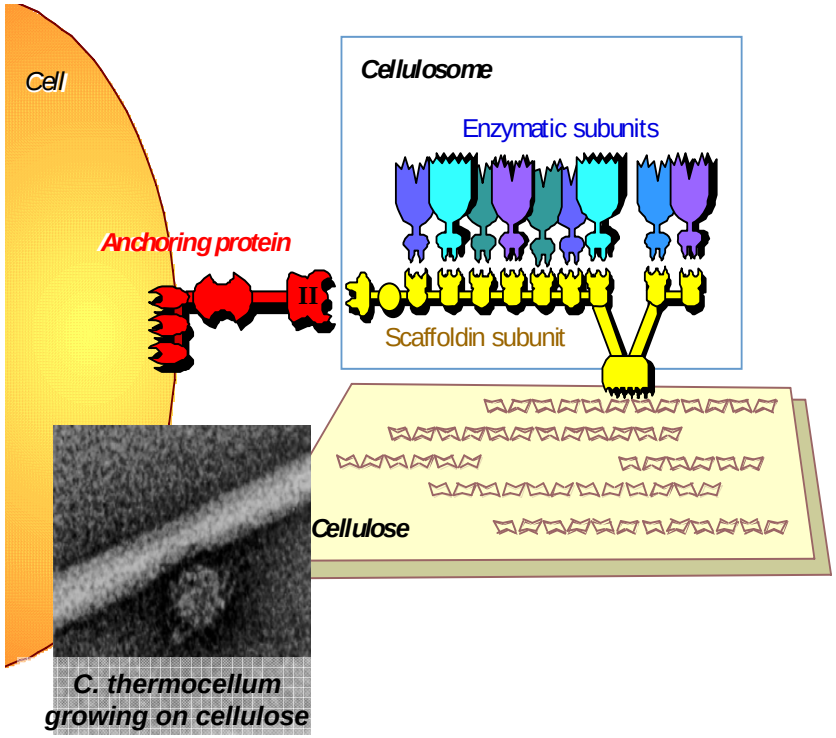


Multi-scale modeling: Cellulose degradation

System of energy research (DOE) interest

- Renewable energy: ethanol production from cellulose
- Detailed understanding of cellulase enzyme mechanisms from multi-scale modeling
  - 1-100 ns trajectories for systems with over 800,000 atoms

- Simulations with different substrates and mutant enzymes
- Creation of more efficient enzymes for cellulose degradation through protein engineering

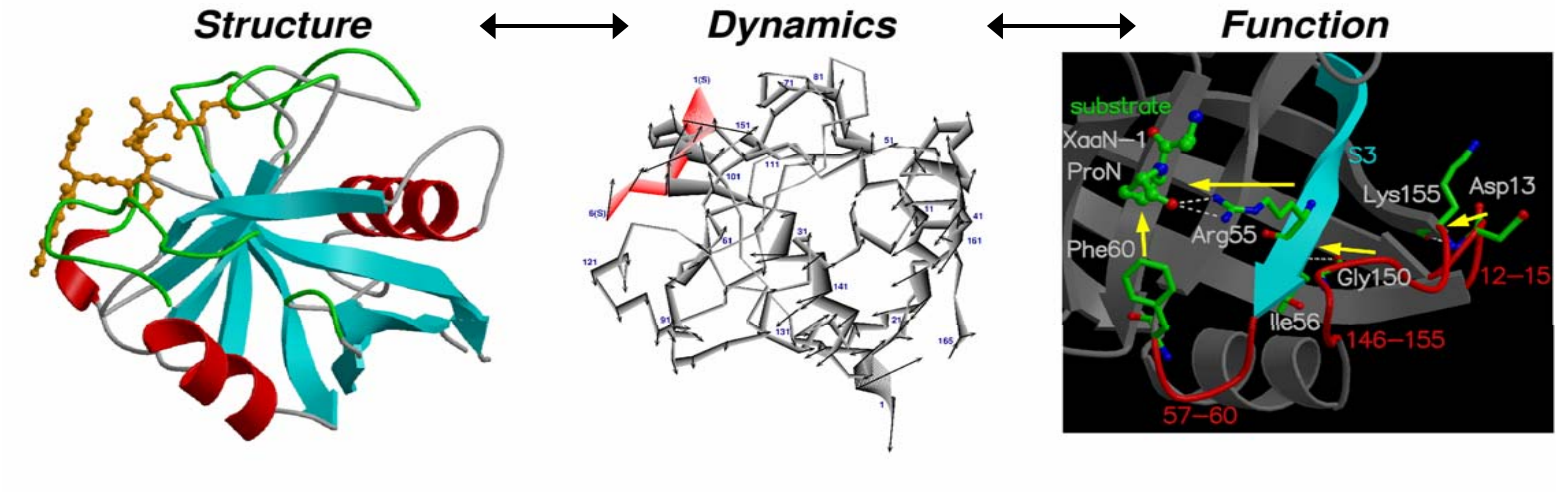


Multi-scale modeling: Enzyme Cyclophilin A

System of medical interest

- Enzyme cyclophilin A plays a role in many cellular function including protein folding and transport
- Required for infectious activity of HIV-1
- Link between enzyme structure, dynamics and function
- Multi-scale modeling lead to identification of

  Network of protein dynamics promoting enzyme catalysis
- Computational finding verified experimentally

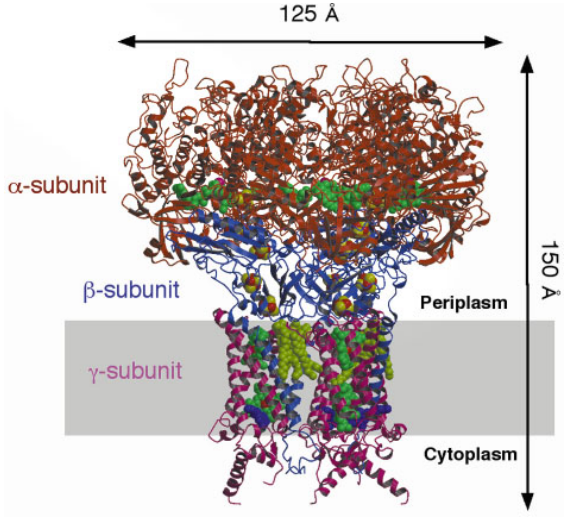


Interfacing simulations with experiments

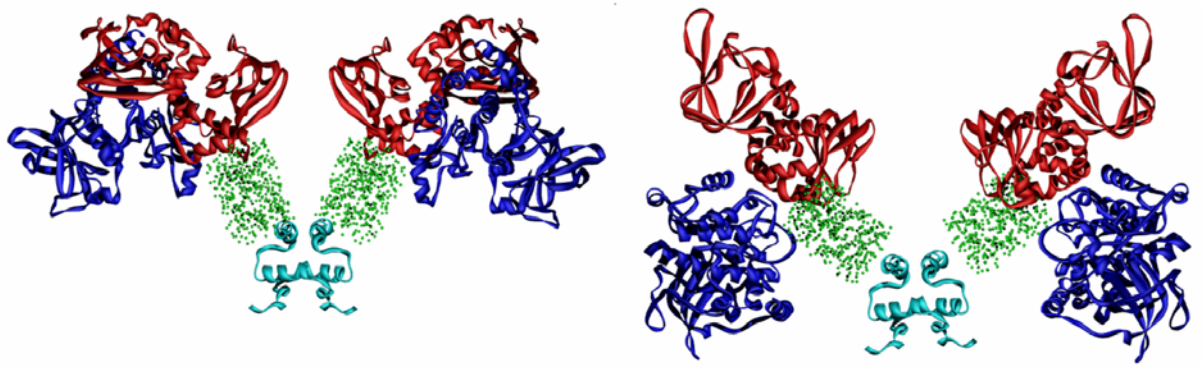
Include data from a variety of experimental techniques into simulations - better models

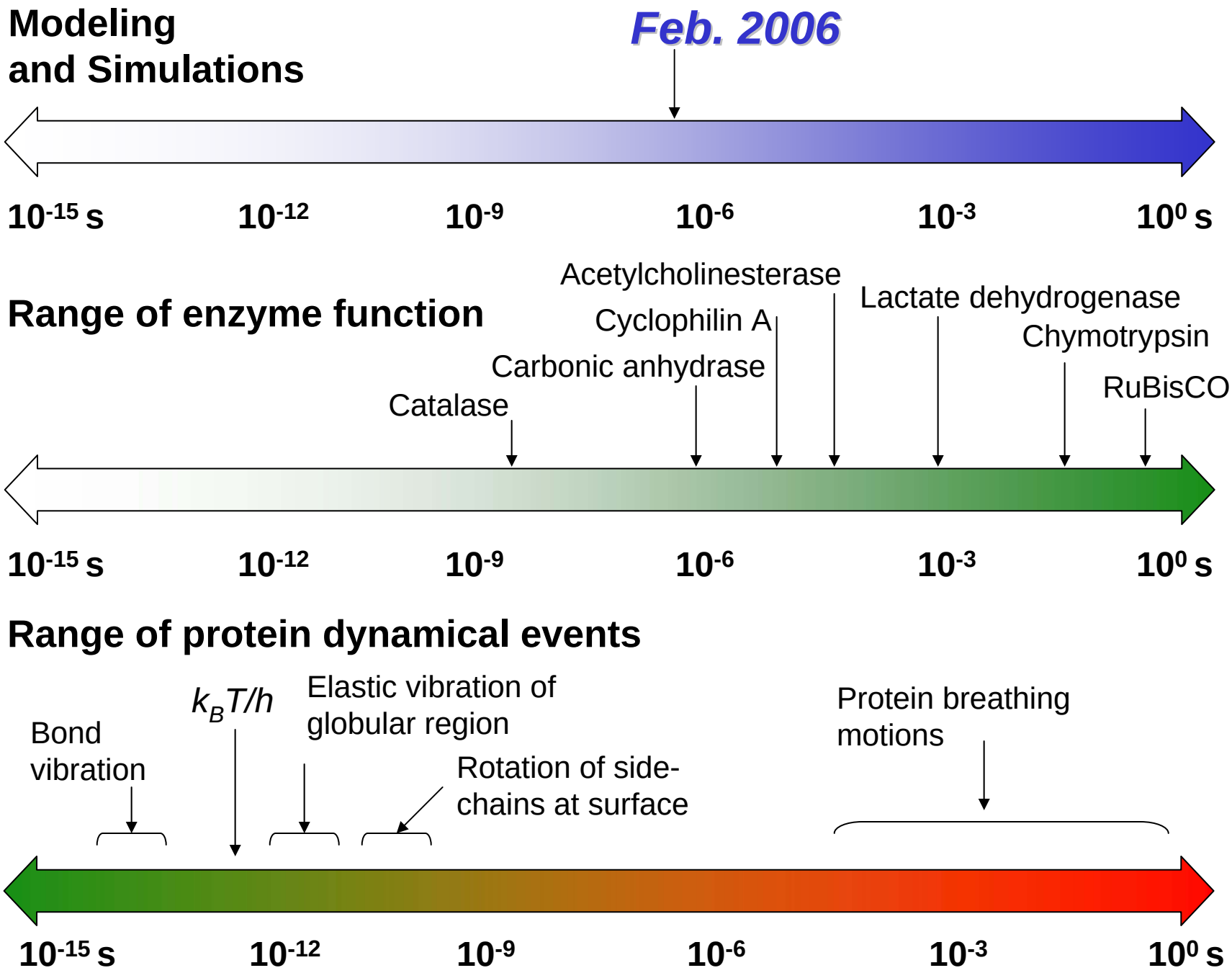
Neutron and X-ray scattering

- RuBisCO
- models of membrane systems
- structure, dynamics and function of biomolecular complexes



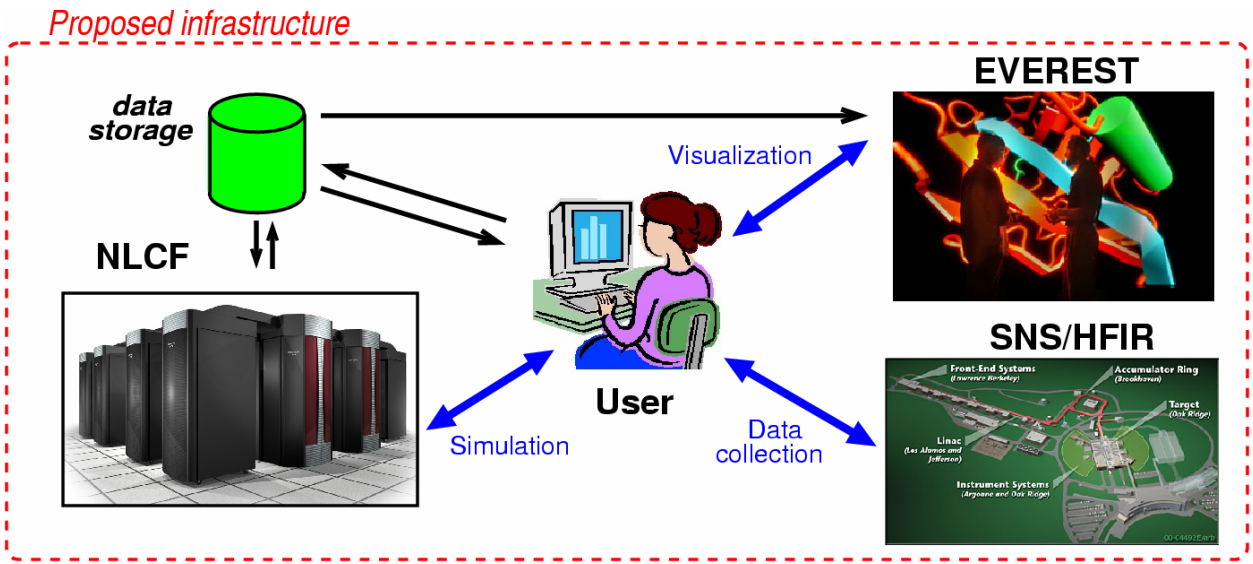
Refining low resolution data through modeling & simulations




Modeling
and Simulations
Feb. 2006
10-15 s
10-12
10-9
10-6
10-3
100 s
Range of enzyme function
Acetylcholinesterase
Cyclophilin A
Lactate dehydrogenase
Chymotrypsin
Carbonic anhydrase
RuBisCO
Catalase
10-15 s
10-12
10-9
10-6
10-3
100 s
Range of protein dynamical events
Bond
vibration
kBT/h
Elastic vibration of
globular region
Rotation of side-
chains at surface
Protein breathing
motions
10-15 s
10-12
10-9
10-6
10-3
100 s

Summary

- Next Generation Modeling and Simulations
- Integrate ORNL's experimental, computing and visualization facilities
- Biophysical characterization of molecular machines through multi-scale modeling
- Impact Chemistry and Materials/Nanotechnology



Acknowledgements

- Thomas Zacharia
- Jeff Nichols
- Al Geist
- Buddy Bland
- DOE
- NCCS